11

Substructure Determination in Isomorphous Replacement and Anomalous Diffraction Experiments

R.W. Grosse-Kunstleve and T. R. Schneider

AU: please provide authors' first names

Summary

The determination of the substructure of heavy atoms or anomalous scatterers is a central step in experimental-phasing procedures. We give an overview of commonly used methods for substructure determination, including estimation of substructure structure factors, Patterson methods, direct methods, dual-space recycling procedures, and methods for substructure refinement and completion. This chapter also includes an annotated list of available program packages.

Key Words: Patterson methods; direct methods; heavy atoms; dual-space recycling; fast translation function.

1. Introduction

Traditionally, experimental phasing of macromolecular structures involves heavy atom soaks and the collection of two or more datasets: the diffraction intensities of the *native* crystal and that of the *derivative(s)*. This is often referred to as single or multiple isomorphous replacement (SIR, MIR). In recent years, because of the growing availability of tunable synchrotron radiation, it has become very popular to use crystals containing anomalous scatterers. Experiments in which the anomalous signal is explicitly measured are known as single or multiple anomalous diffraction experiments (SAD, MAD) ***(1)*** or alternatively single anomalous scattering experiments. Isomorphous replacement and anomalous diffraction can also be combined (SIRAS, MIRAS).

AU: please spell out SIRAS and MIRAS

Heavy atoms, which may at the same time be anomalous scatterers, can be naturally present, for example iron in heme proteins, but more often they are introduced artificially. Extensive overviews of procedures for the preparation of heavy atom derivatives are given by **refs. *2–5***. These procedures are still commonly used. Recently, Dauter ***(6)*** (*see* Chapter 21) introduced the method of

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2*
Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

quick halide soaks. A more complex but powerful and now very popular method for the introduction of anomalous scatterers is the replacement of methionine residues in protein structures by selenomethionine ***(7)***. Interestingly, the data-collection technology developed for selenomethonine experiments has now advanced to the point where even the small anomalous signal from the native sulfur atoms ***(8,9)*** in proteins or phosphorus atoms in oligonucleotides ***(10)*** can in favorable cases lead to successful structure determination. There have also been reports ***(11)*** of a selenomethionine derivative being used together with a native crystal in an isomorphous replacement experiment. This is unusual but entirely possible if the degree of nonisomorphism is sufficiently small.

AU: Please check the chaper x-refs here and throughout. There is no chapter 21; chap 8 meant?

Once crystals containing heavy atoms or anomalous scatterers are available, experimental phasing can be viewed as a divide-and-conquer technique where the larger problem of determining the complete macromolecular structure is divided into two steps:

1. Given experimental diffraction data from one or several crystals, approximate substructure structure factors corresponding to the substructure of heavy atoms or the anomalous scatterers alone are computed. The substructure is solved based on the substructure structure factors using methods originally developed for the solution of small molecules.
2. Using the known substructure in combination with the measured diffraction data, algebraic or probabilistic methods are used to extrapolate phases for the full structure.

Once initial phases are found, the structure solution process continues with density modification, model building, and refinement. In this chapter, we focus on the determination of the substructure and some aspects of substructure refinement.

2. Overview of Commonly Used Methods

Most computer programs in use implement a combination of several procedures. In the following, the most important building blocks are described.

2.1. Estimation of Substructure Structure Factors

2.1.1. Isomorphous Differences

Because the number of atoms in a native macromolecular structure is usually much larger than the number of additional heavy atoms in a derivative, it is a valid approximation to assume $F_H << F_{PH}$, where F_H are the structure factor amplitudes corresponding to the substructure only, and F_{PH} the structure factor amplitudes of the derivative. This approximation leads to (***[2]***; **Subheading 6.2.**):

AU: There is no Subheading 6.2., please modify

$$F_H << F_{PH} \Rightarrow F_{PH} - F_P \approx F_H \cos(\varphi_{PH} - \varphi_H) \quad (1)$$

The cosine term takes on values between –1 and 1. Therefore the isomorphous differences $F_{PH} - F_P$ as extracted from diffraction experiments on the

derivative and the native crystal are lower estimates of the true substructure structure factor modulus F_H: the true F_H can be larger, but they cannot be smaller than the observed isomorphous differences.

2.1.2. Anomalous Differences

AU: There is no Subheading 7.6, please modify

Similar considerations lead to the following equation for anomalous differences $F^+_{PH} - F^-_{PH}$ (***[2]***; **Subheading 7.6.**):

$$F''_H << F'_{PH} \Rightarrow F^+_{PH} - F^-_{PH} \approx 2F''_H \sin(\varphi_{PH} - \varphi_H) \quad (2)$$

Here F''_H are the imaginary contributions to the structure factors of the anomalous scatterers. F'_{PH} is the sum of the structure factors of the macromolecular structure and the real contributions of the anomalous scatterers. The sine term also takes on values between –1 and 1. Therefore, the anomalous differences taken between the measured Bivoet mates F^+_{PH} and F^-_{PH} are lower estimates of the imaginary contributions of the anomalous scatterers.

The ratio between anomalous differences measured at different wavelengths is, to a first approximation, a constant for all reflections from a given crystal. The degree to which experimental data follow this expectation can be measured by calculating standard correlation coefficients (e.g., **ref. *12***) between anomalous differences originating from different measurements. The magnitude of the correlation coefficient is a very useful indicator of data quality (**Fig. 1A**) (*see* **Notes 1–3**).

2.1.3. F_A Structure Factors

In the case of MAD experiments it is possible to compute better estimates of the structure factors corresponding to the substructure. Commonly these estimates are referred to as F_A structure factors. Various algorithms for the computations of F_A structure factors are available: MADSYS ***(1)***, CCP4 REVISE ***(13)***, SOLVE ***(14)***, XPREP (Bruker AXS, Madison, WI). For good MAD data, F_A structure factors usually lead to significantly more efficient determination of the substructure. It has also been observed that F_A structure factors enable the solution of a substructure that could not be solved from anomalous differences measured at any one wavelength. However, if the MAD data are affected by systematic errors, such as intensity changes resulting from radiation damage, it is also possible that the corresponding F_A structure factors are not suitable for the substructure determination. In this case it is advantageous to attempt substructure determination with the wavelength collected first (ideally at the peak of the anomalous signal) ***(15)*** (*see* **Notes 4–8**).

2.2. The Phase Problem

In the second and third decades of the 20th century early X-ray crystallographers worked out that the observed diffraction intensities are directly related to

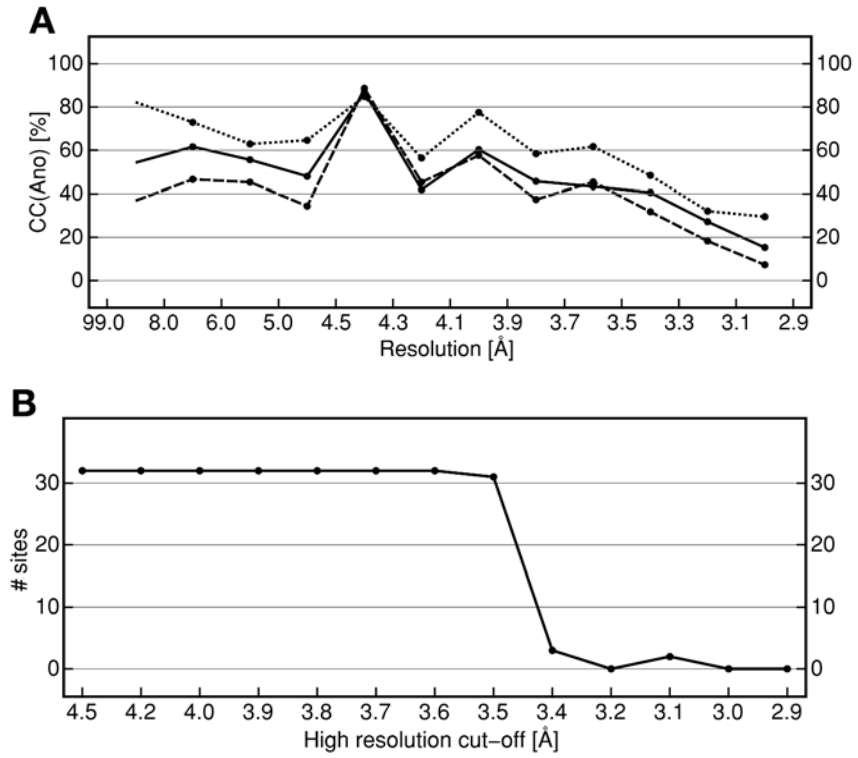


Fig. 1. (A) Plot of the correlation coefficient between signed anomalous differences measured at different wavelengths against resolution for the three-wavelength multiple anomalous diffraction (peak, inflection point, high-energy remote) dataset collected on the Se-Met substituted form of an acyltransferase ***(56)***. The correlations shown are peak-remote (full line), inflection-remote (dashed line), and peak-inflection (dotted line). (B) Plot of the number of correct sites (as checked with SITCOM ***[57]***) present in the top-scoring solution of the Se substructure of the acyltransferase as obtained by SHELXD vs high-resolution cut-off. The crystal structure contains 32 ordered Se-sites.

the Fourier transformation of the electron density of the crystal structure (not taking Lorentz factors, polarization factors, and other experiment-specific corrections into account):

$$\sqrt{I} \equiv |F| \propto FT(\rho) \tag{3}$$

Here I represents the observed intensities, $|F|$ the *structure factor amplitudes*, ρ is the electron density, and FT a Fourier transformation. The same relation more specifically:

$$F_h = |F_h| e^{i\varphi_h} = \frac{V}{N} \sum_x \rho(x) e^{2\pi i h x} \tag{4}$$

Here h is a Miller index, x the coordinate of a grid point in real space, N the total number of grid points, and V the volume of the unit cell. The complex structure factor F is also shown in the alternative representation as an amplitude $|F|$ combined with a phase (ϕ).

Obviously it is straightforward to compute the structure factor amplitudes from the electron density. Given complex structure factors it is equally easy to compute the electron density via a Fourier transformation:

$$\rho \propto FT^{-1}(F) \tag{5}$$

where FT^{-1} represents the inverse Fourier transformation. Unfortunately with current technology it is almost always impractical to directly measure both intensities and phases. Conventional diffraction experiments only produce intensities, the phases are not available. This is colloquially known as the phase problem of crystallography.

2.3. Techniques for Solving the Phase Problem

2.3.1. Patterson Methods

The Patterson function is defined as the Fourier transformation of the observed intensities:

$$Patterson \propto FT^{-1}(I) \tag{6}$$

This is a straightforward calculation requiring only the experimental observations as the input. Patterson ***(16)*** showed that the peaks in this Fourier synthesis correspond to *vectors between atoms* in the crystal structure.

2.3.2. Patterson Interpretation in Direct Space and in Reciprocal Space

In the classic textbook *Vector Space* Buerger ***(17)*** demonstrates that under idealized conditions image seeking procedures are capable of recovering the image of the electron density from the Patterson function. "Idealized conditions" essentially means fully resolved peaks in the Patterson function. In practice, this condition is only fulfilled for very small structures, but it still is possible to extract useful information from real Patterson maps. The basic idea is:

1. Postulate a hypothesis, for example a putative substructure configuration.
2. Test the hypothesis against the Patterson map.

The test involves the computation of vectors between the atoms of the putative substructure and the determination of the values in the Patterson map at the location of these vectors. This involves interpolation between grid points of the

map. The interpolated peak heights are usually the input for the computation of a Patterson score. Theoretically the minimum of all the peak heights found is the most powerful measure, but sum or product functions have also been used ***(17)***. Nordman ***(18)*** suggests using the mean of a certain percentage of the lowest values.

It is also possible to work with the observed intensities in reciprocal space, without transforming them according to **Eq. 6**. Conceptually the procedure is even simpler:

1. Postulate a hypothesis, for example a putative substructure configuration.
2. Test the hypothesis against the observed intensities.

In this case the test involves the calculation of intensities for the putative structure and the evaluation of a function comparing these with the observed intensities, for example the standard linear correlation coefficient (e.g., **ref. *12***). An advantage of this method is that it does not involve interpolations and therefore it should be intrinsically more accurate. However, the calculations are much slower than the computation of Patterson scores in direct space if done in the straightforward fashion suggested here. The key to making the reciprocal space approach feasible is the fast translation function devised by Navaza and Vernoslova ***(19)***. We were able to show that the fast translation function is typically 200–500 times faster than the conventional translation function. The fast translation function was originally designed for solving molecular replacement problems, but it is also being used for the determination of substructures ***(20,21)***.

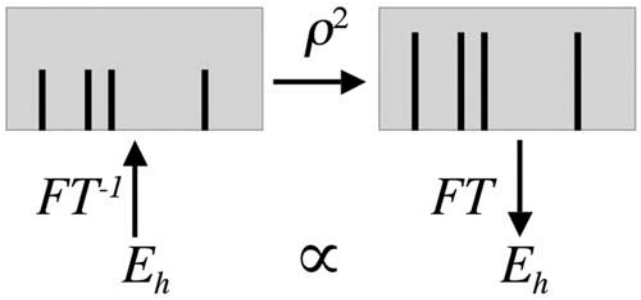
2.3.3. Direct Methods

Direct methods were originally developed for the direct determination of phases without direct use of stereochemical knowledge. The fundamental approach is to start with a very small set of starting phases and to construct a more complete phase set by applying phase probability relationships. The expanded phase set in combination with the observed structure factors is used to compute an electron density map that is hopefully interpretable when stereochemical knowledge is taken into account.

The phase probability relations governing the phase extension procedure are usually based on the well-known tangent formula ***(22)***. This formula is typically introduced as:

$$\tan(\varphi_h) = \frac{\sum_k |E_k E_{h-k}| \cos(\varphi_k + \varphi_{h-k})}{\sum_k |E_k E_{h-k}| \sin(\varphi_k + \varphi_{h-k})} \tag{7}$$

The tangent formula can be used to estimate an unknown phase φ_h given a set of known phases φ_k and φ_{h-k}. To avoid distraction, for the moment we will assume that the *E*s in this formula are analogous the structure factors *F* previ-



AU: reverse paragraph symbol removed for Fig. 2. Ok?

Fig. 2. The essence of direct methods. Normalized structure factors correspond to point atoms at rest. Squaring in direct space followed by a Fourier transformation leads to structure factors that are proportional to the original structure factors. The phases are identical.

ously introduced (the exact definition of the *E*s is given by **Eq. 11**). The derivation of the tangent formula employs the assumptions that the electron density is positive everywhere in the unit cell (positivity) and that all atoms are resolved (atomicity). To understand this it is useful to rewrite the tangent formula as a simpler, but mathematically equivalent, expression:

$$E_h \propto \sum_k E_k E_{h-k} \tag{8}$$

Comparison with the definition of the convolution (e.g., **ref. *23***) leads us to recognize:

$$\sum_k E_k E_{h-k} \equiv Convolution(E, E) \tag{9}$$

Application of the convolution theorem followed by application of **Eq. 5** leads to:

$$E_h \propto \sum_k E_k E_{h-k} = FT(FT^{-1}(E)^2) = FT(\rho^2) \tag{10}$$

This equation shows that the tangent formula uses positivity and atomicity to introduce a self-consistency argument as illustrated in **Fig. 2**. This argument is essentially the same one used in the derivation of the Sayre equation ***(24)***, which is slightly more complex than the tangent formula because it is formulated for atoms with Gaussian shapes rather than point atoms. This describes real crystal structures more closely, but in practice it is often more advantageous to eliminate the shape term and to work with normalized structure factors corresponding to point atoms. Normalized structure factors can be estimated from observed structure factor amplitudes $|F_h|$ by enforcing the expected average $\left\langle |E_h|^2 \right\rangle = 1$ in resolution shells:

$$|E_h|^2 = \frac{|F_h|^2}{\left\langle |F_h|^2 \,/\, \varepsilon_h \right\rangle} \tag{11}$$

To be precise, this equation yields estimates of the quasi-normalized structure factors. The term ε takes the multiplicity of the reflections into account and can be directly computed from the space group symmetry.

2.3.4. Dual-Space Recycling

The tangent formula alone often does not work efficiently for solving structures with many atoms ***(25)***. The most popular "direct methods" programs used in macromolecular crystallography today are the result of an evolution that transformed the pure phase-extension idea into complex multitrial search procedures. MULTAN ***(26)*** pioneered the multitrial approach but is still motivated by the phase-extension idea. RANTAN ***(27)*** and early versions of SHELX ***(28)*** mark the transition to random-seeded multitrial approaches that use the tangent formula in a recycling procedure to enforce self-consistency (**Fig. 2**). Shake-and-Bake ***(29)*** and more recent versions of SHELX ***(30)*** introduced the concept of dual-space recycling ***(31)***. Reciprocal-space phase manipulation based on the tangent formula, or the minimal function in the case of shake-and-bake, alternated with direct-space interpretation of Fourier maps. Shake-and-Bake picks peaks from the Fourier maps, taking a given minimum distance into account. The peaks are used in a structure–factor calculation to obtain new phases that are entered into the next cycle of phase manipulation. SHELXD ***(32)*** follows a similar approach but after selecting typically 1.3 times the expected number of sites N, eliminates 0.3 N of them in a random omit procedure. This imitates the common practice in macromolecular refinement of omitting part of a structure from a Fourier calculation in order to obtain an improved electron density in the next iteration. Picking peaks from an electron density to be used for the calculation of the next phase set is a very direct way of enforcing atomicity and taking into account the actual number of scatterers. The dual-space recycling procedure is very fast as it only works with a small fraction of the reflections. Typically 10–15% of the strongest E-values are used.

2.3.5. Direct Methods Recycling With Patterson Seeding

Conventional direct methods programs initialize the recycling procedure with random phases or random coordinates. In contrast, SHELXD ***(32)*** and HySS ***(21)*** use Patterson seeding to obtain better than random starting phases for the recycling procedure. The fundamental steps in the procedure are:

1. Generation of two-atom fragments. A given number of peaks are picked from a sharpened Patterson map. These are considered to be possible vectors between two

atoms of the substructure. However, at this stage only the relative orientation of the two atoms is known, but not the absolute position in the unit cell. SHELXD uses the Nordman ***(18)*** function to obtain scores for a large number of random translations of the two-atom fragments. HySS uses the fast translation function ***(19)*** to systematically score all translations on a uniform grid.

2. Extrapolation to the full substructure. Conceptually a third probe atom is systematically placed on the points of a uniform grid over the asymmetric unit while keeping a trial two-atom fragment fixed at a position that lead to a high score. SHELXD uses the Nordman function to compute scores for the position of the third atom. HySS uses the fast translation function for the same purpose. Points with the highest scores are added to the original two-atom fragment to generate the expected number of atoms.
3. Correction of defects. Typically the structures obtained in the previous step contain a considerable number of misplaced atoms. Even the best solutions often have less than half of the atoms correctly placed. These defects are efficiently corrected using dual-space recycling (tangent formula expansions and random omission of peaks). The standard linear correlation coefficient between calculated and observed intensities is a very reliable score for ranking the final results of the dual-space recycling procedure.

2.4. Identification of Correct Solutions

Most search procedures including shake-and-bake and SHELXD are run for a preset number of trials, or until they are terminated manually. Various figures of merit, such as the Patterson figure of merit to measure the consistency between calculated and observed Patterson maps ***(33)***, the value of the minimal function for the final set of phases ***(29)***, or the correlation coefficient between observed and calculated E-values, can then be used to compare the results of different trials. However, a general problem with these figures of merit is that an absolute threshold for the discrimination between correct and incorrect solution cannot be given. Some searches yield trimodal distributions so that simply looking for outstanding figures of merit can also be misleading. Thus, additional criteria have to be employed. One such criterion is the presence of non-crystallographic symmetry in the substructure. This can be checked for example by using the "Patterson crossword table" ***(33)*** as implemented in SHELXD, by using a sophisticated combination of scores ***(34)*** or simply by inspecting the substructure on a graphical display. Another useful criterion for the correctness of a substructure that is suitable for automation is the consistency of substructures obtained from different trials. Therefore, the shake-and-bake suite ***(35)*** and recently the SHELX suite ***(36)*** include programs for comparing substructures that automatically take allowed origin shifts and change-of-hand operators into account. These programs are run externally from the actual search procedure. In HySS a similar procedure is fully integrated into the substructure

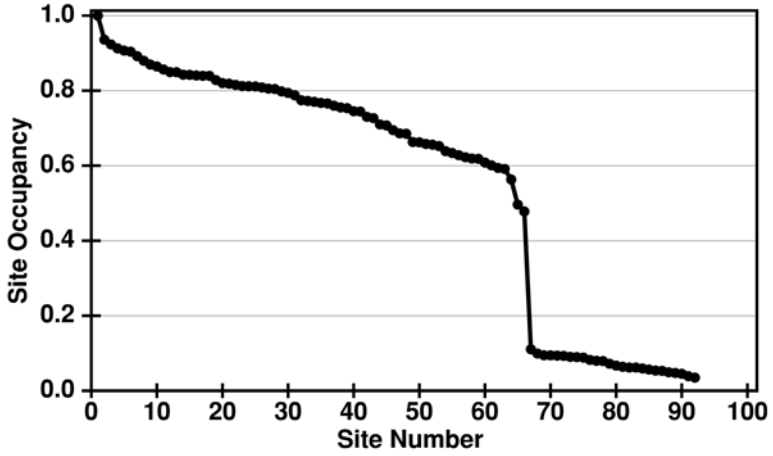


Fig. 3. Occupancy vs site number for a typical solution obtained for the Se-substructure of 2-aminoethylphosphonate transaminase ***(58)*** against FA-data produced by XPREP. The occupancy of site 66 is 0.48 and the occupancy of site 67 is 0.11, clearly indicating that the substructure contains 66 sites.

search and is used in combination with the correlation coefficients to automatically terminate a search ***(21)***.

Often it is not clear how many well-ordered substructure sites to expect exactly, in particular when working with heavy atom soaks (*see* **Notes 9** and **10**). It is also quite common that a certain fraction of selenomethionine residues is not well ordered, for example residues near the termini. The SHELXD program employs refinement of occupancy factors to assist in determining the number of ordered sites. At the end of each trial a certain number of the highest peaks, typically 1.3 times the number of expected sites, are picked from the last E-map obtained in the dual-space recycling procedure. The occupancies of these sites are then refined against all observed substructure structure factors and the sites are sorted according to the refined values of their occupancies. In favorable cases the plot of the occupancy vs site number shows a sharp drop, indicating where the list of well-ordered sites ends (**Fig. 3**). Only sites that appear before this point should be used in subsequent steps.

Occupancy refinement is also used in the HySS program, but only after the determination of the sites common to the top two or three solutions (the consensus model), ranked by the correlation coefficient. The consensus model may contain more sites than expected because the number of peaks picked from the last E-map depends on the number of expected sites N. Typically, for $N < 30$ between N and 1.5 N peaks are picked. All these sites are subject to positional and occupancy refinement, and the resulting model sorted by occupancies in

decreasing order. Only the top *N* sites are retained. If this means that the model was truncated, it is refined a second time. Experience shows that the refined, truncated consensus model rarely contains a significant number of incorrect sites, if any.

2.5. Substructure Refinement and Completion

All programs currently in use for the determination of substructures will produce a model for the substructure that has the best agreement with a given set of substructure structure factors. However, for the determination of the best phases on which the initial electron density map for the complete structure will be based, it may be advantageous to use a substructure model that is consistent with all experimental data. For instance, in MAD phasing, the substructure should not only be in agreement with the the F_A-values, but more importantly should be consistent with the two or more original datasets collected at different wavelengths.

The process of refinement and completion of a substructure model against all available data in many aspects resembles the refinement of an entire macromolecular against native data: parameters of the model are optimized such that the agreement with the experimental data is improved and as the model becomes better, difference electron density maps reveal previously invisible parts of the structure to be included into the model. There are, however, a number of aspects that make substructure refinement substantially more complex than "native" structure refinement. As substructures contain significantly fewer atoms than structures of entire proteins, the data-to-parameter ratio is much higher, supporting the refinement of more parameters per site. Commonly refined parameters not only include coordinates and B-values, but also anisotropic B-values, occupancies, and dispersive and anomalous contributions to the scattering factor. Some of these parameters are highly correlated, e.g., the B-value and the occupancy of a site, and appropriate constraints and restraints have to be introduced to obtain reasonable results. Furthermore, the parameters have to be adjusted such that the agreement with several datasets is improved at the same time. This simultaneous use of several datasets introduces the problem of the relative scaling and weighting of the data from different sources.

When datasets from different crystals (or from the same crystals with different levels of radiation damage) are used, a further complication is introduced by the fact that structural differences between the different crystals, the so-called nonisomorphism, result in inconsistencies ("lack-of-closure") in the phasing calculations that are not trivial to handle. The difficulty of relative weighting of different sources of information also becomes apparent in the calculation of intermediate electron density maps to be used for completing the substructure model and a variety of sophisticated schemes have been suggested, one prominent example being the log-likelihood gradient map ***(37)***.

Several programs, mostly based on maximum-likelihood principles, are available for carrying out heavy atom refinement and phasing. These include CNS ***(38)***, MLPHARE ***(39,40)***, PHASES ***(41)***, SHARP ***(37)***, and SOLVE ***(34)***.

With currently available experimental technology, data of very high quality can be collected (*see* **Note 11**) and in favorable cases it can be afforded to be less strict in the treatment of experimental errors for the sake of simplicity and speed. The recently implemented combination of SHELXD and SHELXE follows this route: based on the unmodified substructure model found by SHELXD, SHELXE rapidly calculates an initial phase set that is subsequently subjected to density modification, often yielding an interpretable electron density map within minutes.

3. Summary of Programs Available

ACORN: http://www.ccp4.ac.uk/. Primarily designed for solving protein structures at atomic resolution (1.2 Å or better), but tests of substructure determinations were also reported ***(42)***. Starts with a known substructure, known fragments positioned with an extensive molecular replacement search, or randomly placed atoms. This is followed by application of Patterson superposition methods, Sayre phase refinement, and a special density modification procedure (dynamic density modification).

CNS heavy_search.inp: http://cns.csb.yale.edu/. Evaluation of direct-space and reciprocal-space Patterson interpretation, combined with Patterson correlation refinement. A robust but relatively slow method that has proven most useful for solving moderately sized structures from noisy data ***(43)***, but it has also been tested successfully to solve structures with up to 66 sites ***(44)***.

AU: see in where below. Give exact location.

HySS: http://phenix-online.org/. Dual-space recycling with Patterson seeding. A highly automated procedure with a sophisticated, fully integrated validation mechanism that enables automatic termination as soon as the solution is clear. Particularly easy to use given anomalous diffraction data because no input files other than the reflection file are required and all common reflection file formats are processed directly (this includes files produced by XPREP, *see* below).

AU: there is no chapter 25, please clarify.

RANTAN: http://www.ccp4.ac.uk/. Tangent refinement initialized with random phases. Used by autoSHARP (http://www.globalphasing.com/sharp/) (*see* Chapter 25).

Shake-and-Bake: http://www.hwi.buffalo.edu/SnB/. Dual-space recycling, alternatively using the tangent formula or the minimal function ***(45)***. The complete shake-and-bake suite includes sophisticated procedures for the determination of normalized difference structure factors ***(46)***, graphical visualization tools for monitoring the progress of the search, and a procedure for substructure comparisons ***(35)***.

SHELX: http://shelx.uni-ac.gwdg.de/. The SHELX-suite of programs for macromolecular phasing includes modules for the determination of substructure structure factors (SHELXC), substructure solution (SHELXD), phase calculation, and density modification (SHELXE). The modules can be used individually or in an integrated manner via a graphical user interface (HKL2MAP).

SOLVE & RESOLVE: http://solve.lanl.gov/. SOLVE implements a tight integration of direct-space Patterson methods, difference Fourier analysis, and phasing ***(34)***. One or two initial substructure sites are determined with Patterson superposition functions. The remaining sites are found by repeated analysis of isomorphous or anomalous difference Fourier maps. These fundamental building blocks are integrated into a high-level procedure that automates decision-making using a sophisticated scoring system. Includes all steps including the refinement of experimental phases. The RESOLVE program ***(47)*** implements statistical density modification and automated model building.

XPREP: http://www.bruker-axs.com/. Program for the evaluation of diffraction data. Among other features, it allows the estimation of substructure structures from various types of experiments, such as MAD, SAD, MIR, SIRAS, and others. The program contains procedures for detwinning anomalous data.

4. Notes

1. Careful evaluation of data quality is crucial. The structure factors representing a substructure are calculated as small differences between large experimentally measured values. Therefore, the importance of evaluating the data quality before starting the substructure solution process cannot be overstated. If measurements of anomalous differences at different wavelengths are available, truncating the data at the resolution where the correlation coefficient between anomalous differences falls below 30% is a good strategy (**Fig. 1A**). Including data of lower quality may be harmful and in extreme cases the inclusion of a 0.1-Å thick high-resolution shell can lead to failure of the substructure solution process (**Fig. 1B**). Another measure of data quality is the signal to noise for the substructure factors. However, for this measure absolute thresholds are difficult to define because the estimation of accurate standard deviations for substructure structure factors is difficult.

   In a single-wavelength experiment it may be advantageous to carry out additional steps in order to obtain two or more independent measurements of anomalous differences. For example, the crystal can be rotated around different axes on a multicircle diffractometer. A very rigorous approach is to repeat the same experiment on two different crystals and then to correlate the measurements (e.g., **ref. *48***).
2. Some programs provide the facilities to use unmerged data (e.g., SOLVE and XPREP). It is often advantageous to provide unmerged data to these programs. In comparison to premerged data, unmerged data allow for a more robust statistical analysis. A major reason is that uncertainties can be estimated from a spread of measurements and outliers can be detected more reliably.

3. Collecting highly redundant data has been repeatedly shown to increase the chances of success ***(10,49)***. In addition, higher redundancy results in more reliable statistics.
4. Radiation damage is the major obstacle to collecting data to high redundancy. If the crystal is heavily affected by radiation damage, a low-dose, high-redundancy dataset that extends to 3 Å can be much more useful for initial phasing than a high-dose, low-redundancy dataset that extends to 2 Å.
5. Radiation damage can be used as a phasing tool. The specific structural changes induced in a crystal structure when deliberately exposed to large doses of X-ray photons can be exploited for experimental phasing. Ravelli et al. ***(50)*** have successfully used a dataset from a fresh crystal as the native dataset, and the dataset from a "burnt" crystal as the derivative dataset in a pseudo isomorphous replacement experiment.
6. Collecting related reflections close in space and time (*see* Chapter 18) can make the difference between success and failure if radiation damage is a severe problem (e.g., **ref. *51***).
7. Merohedral twinning ***(52)*** creates fewer problems than one would expect for the determination of the substructure as: (1) the strong reflections used in dual-space recycling procedures are statistically less affected by the twinning than general reflections, and (2) structure solution using Patterson methods against twinned data will produce only one twin component as no vectors between the twin component are present in the Patterson map ***(53)***.
8. Treating a failed MAD experiment as a SAD experiment may be a viable approach. Sometimes, the systematic errors between the different wavelengths of a MAD experiment are too large to give useful results. Using selected single-wavelength data in a SAD approach may nonetheless solve the structure ***(43)***.
9. The simultaneous presence of different kinds of anomalous scatterers can lead to complications because the wavelength-dependent change of the anomalous signal is very specific to each element type. The available substructure determination programs may not function properly because a number of commonly used approximations are based on the assumption that only one kind of anomalous scatterer is present.
10. If the number of substructure sites is not known, e.g., in the case of halide soaks (*see* Chapter 21), it is generally better not to overestimate the expected number sites, but to err on the low side. If too many spurious sites are included in the calculations most figures of merit cease to be good indicators. A rule of thumb could be "less than number of residues divided by 20," but often multiple different values have to be tried.
11. MAD phasing at atomic resolution is possible. Recent work on endoglucanase A ***(54)*** and aldose reductase ***(55)*** with experimental phases extending to 1.0 and 0.9 Å resolution, respectively, has shown that a well-refined heavy atom substructure can provide the basis for the calculation of excellent phases to atomic resolution. In favorable cases the maps are virtually indistinguishable from refined maps.

AU: Please provide correct chapter x-ref in 6 and 10

Acknowledgments

We thank P. Adams for suggestions and for critically reading this manuscript. Work by RWGK was funded in part by the US Department of Energy under contract no. DE-AC03-76SF00098 and by National Institutes of Health

(NIH)/NIGMS under grant number 1P01GM063210. Work by TRS was funded in part by the European Union (QLRI-CT-2000-00398).

References

1. Hendrickson, W. A. (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Sci*e**nce 254,** 51–58**.**
2. Blundell, T. L. and Johnson, L. N. (1976) Protein Crystallography. Academic Press, London, UK.
3. McPherson, A. J. (1982) The Preparation and Analysis of Protein Crystals. Wiley, NY.
4. Petsko, G. A. (1985) Preparation of isomorphous heavy-atom derivatives. *Meth. Enzym.* **114,** 147–156.
5. Carvin, D., Islam, S. A., Sternberg, M. J. E., and Blundell, T. L. (2001) The preparation of heavy atom derivatives of protein crystals for use in multiple isomorphous replacement. In: *International Tables for Crystallography: Crystallography of Biological Macromolecules, Vol. F,* (Rossman, M. G. and Arnold, E., eds.), Kluwer Academic Publishers, Dordrecht, Germany, pp. 247–255.
6. Dauter, Z. (2002) New approaches to high-throughput phasing. *Curr. Opin. Struct. Biol.* **12,** 674–678.
7. Doublié, S. (1997) Preparation of selenomethionyl proteins for phase fetermination. *Meth. Enzym.* **276,** 523–530.
8. Hendrickson, W. A. and Teeter, M. M. (1981) Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. *Nature (London)* **290,** 107–113.
9. Dauter, Z., Dauter, M., De La Fortelle, E., Bricogne, G., and Sheldrick, G. M. (1999) Can Anomalous signal of sulfur become a tool for solving protein crystal structures? *J. Mol. Biol.* **289,** 83–92.
10. Dauter, Z. and Adamiak, D. A. (2001) Anomalous signal of phosphorus used for phasing DNA oligomer: importance of data redundancy. *Acta Cryst.* **D57,** 990–995.
11. Lemke, C. T., Smith, G. D., and Howell, P. L. (2002) S-SAD, Se-SAD and S/Se-SIRAS using Cu Ka radiation: why wait for synchrotron time? *Acta Cryst.* **D58,** 2096–2101.
12. Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986) Numerical Recipes. The Art of Scientific Computing. Cambridge University Press, Cambridge, UK.
13. Fan, H. F., Woolfson, M. M., and Yao, J. X. (1993) New techniques of applying multi-wavelength anomalous scattering data. *Proc. R. Soc. Lond*. A442, 13–32.
14. Terwilliger, T. C. (1994) MAD phasing: Bayesian estimates of FA. *Acta Cryst.* D50, 11–16.
15. Dauter, Z. (2002) One-and-a-half wavelength approach. *Acta Cryst*. 1958–1967.
16. Patterson, A. L. (1935) A direct method for the determination of the components of interatomic distances in crystals. *Z. Krist*. 90, 517–542.
17. Buerger, M. J. (1959) Vector Space, Wiley, NY.

AU: Please provide city of publisher for ref. 3

AU: Please provide city of publisher for ref. 17

18. Nordman, C. E. (1966) Vector space search and refinement procedures. *Trans. Am. Crystallogr. Assoc.* **2,** 29–38.
19. Navaza, J. and Vernoslova, E. (1995) On the fast translation functions for molecular replacement. *Acta Cryst.* **A51,** 445–449.
20. Grosse-Kunstleve, R. W. and Brunger, A. T. (1999) A highly automated heavy-atom search procedure for macromolecular structures. *Acta Cryst.* **D55,** 1568–1577.
21. Grosse-Kunstleve, R. W. and Adams, P. D. (2003) Substructure search procedures for macromolecular structures. *Acta Cryst.* **D59,** 1966–1973.
22. Karle, J. and Hauptman, H. A. (1956) A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, 3P12, 3P22. *Acta Cryst.* 9**,** 635–651.
23. Giacovazzo, C. (1992) *Fundamentals of Crystallography.* IUCr/Oxford University Press, Oxford, UK.
24. Sayre, D. (1952) The squaring method: a new method for phase determination. *Acta Cryst.* **5,** 60–65.
25. Woolfson, M. M. (1961) *Direct Methods in Crystallography.* Oxford University Press, Oxford, UK.
26. Germain, G., Main, P., and Woolfson, M. M. (1970) On the application of phase relationships to complex structures. II. Getting a good start. *Acta Cryst.* **B26,** 274–285.

AU: is Gottingen the correct city?

27. Yao, J. X. (1981) On the application of phase relationships to complex structures XVIII. RANTAN-random MULTAN. *Acta Cryst.* **A37,** 642–644.
28. Sheldrick, G. M. (1985) *SHELXS86. Program for the Solution of Crystal Structures.* University of Göttingen, Göttingen, Germany.
29. Miller, R., Gallo, S. M., Khalak, H. G., and Weeks, C. M. (1994) SnB: crystal structure determination via shake-and-bake. *J. Appl. Cryst.* **27,** 613–621.
30. Sheldrick, G. M. and Gould, R. O. (1995) Structure solution by iterative peaklist optimization and tangent expansion in space group P1. *Acta Cryst.* **B51,** 423–431.
31. Sheldrick, G. M., Hauptman, H.A., Weeks, C.M., Miller, R., and Uson, I. (2001) Ab initio phasing. In: *International Tables for Crystallography. Crystallography of Biological Macromolecules, Vol. F,* (Rossman, M. G. and Arnold, E., eds.), Kluwer Academic Publishers, Dordrecht, Germany, pp. 233–245.
32. Schneider, T. R. and Sheldrick, G. M. (2002) Substructure solution with SHELXD. *Acta Cryst.* **D58,** 1772–1779.
33. Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H., and Sieker, L. C. (1993) The application of direct methods and Patterson interpretation to high-resolution native protein data. *Acta Cryst.* **D49,** 18–23.
34. Terwilliger, T. C. and Berendzen, J. (1999) Automated MAD and MIR structure solution. *Acta Cryst.* **D55,** 849–861.
35. Smith, G. D. (2002) Matching selenium-atom peak positions with a different hand or origin. *J. Appl. Cryst.* **35,** 368–370.
36. Dall'Antonia, F., Baker, P. J., and Schneider, T. R. (2003) Optimization of selenium substructures as obtained from SHELXD. *Acta Cryst.* **D59,** 1987–1994.

37. De La Fortelle, E. and Bricogne, G. (1997) Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Meth. Enzym.* **276,** 472–494.
38. Brunger, A. T., Adams, P. D., Clore, G. M., et al. (1998) Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Cryst.* **D54,** 905–921.
39. Collaborative Computational Project, N. (1994) The CCP4 suite: programs for protein crystallography. *Acta Cryst.* **D50,** 760–763.
40. Otwinowski, Z. (1991) Maximum likelihood refinement of heavy atom parameters. In: *Isomorphous Replacement and Anomalous Scattering, Proc. Daresbury Study Weekend,* SERC Daresbury Laboratory, Warrington, UK, pp. 80–85.
41. Furey, W. and Swaminathan, S. (1997) PHASES-95: a program package for processing and analyzing diffraction data from macromolecules. *Meth. Enzym.* **277,** 590–620.
42. Yao, J. X. (2002) ACORN in CCP4 and its applications. *Acta Cryst.* **D58,** 1941–1947.
43. Rice, L. M. and Brunger, A. T. (1999) Crystal structure of the vesicular transport protein sec17: implications for snap-mediated snare complex disassembly. *Mol. Cell* **4,** 85–96.
44. Weeks, C. M., Adams, P. D., Berendzen, J., et al. (2003) Automatic solution of heavy-atom substructures. *Meth. Enzym.* **374,** 37–83.
45. DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R., and Hauptman, H. A. (1994) Structure solution by minimal function phase refinement and Fourier filtering: theoretical basis. *Acta Cryst.* **A50,** 203–210.
46. Blessing, R. H. and Smith, G. D. (1999) Difference structure factor normalization for heavy-atom or anomalous-scattering substructure determinations. *J. Appl. Cryst.* **32,** 664–670.
47. Terwilliger, T. C. (2000) Maximum-likelihood density modification. *Acta Cryst.* **D56,** 965–972.
48. Cordell, S. C., Anderson, R. E., and Löwe, J. (2001) Crystal structure of the bacterial cell division inhibitor MinC. *Embo J.* **20,** 2454–2461.
49. Debreczeni, J. E., Bunkóczi, G., Ma, Q., Balser, H., and Sheldrick, G. M. (2003) In-house measurement of the sulfur anomalous signal and its use for phasing. *Acta Cryst.* **D59,** 688–696.
50. Ravelli, R. B., Schroeder Leiros, H. K., Pan, B., Caffrey, B., and McSweeney, S. (2003) Specific radiation damage can be used to solve macromolecular crystal structures. *Structure* **11,** 217–224.
51. Clemons, W. M. J., Brodersen, D. E., McCutcheon, J. P., et al. (2001) Crystal structure of the 30S ribosomal subunit from Thermus thermophilus: purification, crystallization and structure determination. *J. Mol. Biol.* **310,** 827–843.
52. Yeates, T. O. (1997) Detecting and overcoming crystal twinning. *Meth. Enzym.* **276,** 344–360.
53. Rudolph, M. G., Kelker, M. S., Schneider, T. R., et al. (2003) Use of multiple anomalous dispersion to phase highly merohedrally twinned crystals of interleukin-1b. *Acta Cryst.* **D59,** 290–298.

54. Schmidt, A., Gonzalez, A., Morris, R. J., Costabel, M., Alzari, P. M., and Lamzin, V. S. (2002) Advantages of high-resolution phasing: MAD to atomic resolution. *Acta Cryst.* **D58,** 1433–1441.
55. Podjarny, A., Schneider, T. R., Cachau, R. E., and Joachimiak, A. (2003) Structural information content at high resolution: MAD versus native. *Meth. Enzym.* **374,** 321–341.
56. Hensgens, C. M., Kroezinga, E. A., van Montfort, B. A., van der Laan, J. M., Sutherland, J. D., and Dijkstra, B. W. P. (2002) Purification, crystallization and preliminary X-ray diffraction of Cys103Ala acyl coenzyme A:isopenicillin N acyltransferase from Penicillium chrysogenum. *Acta Cryst.* **D58,** 716–718.
57. Dall'Antonia, F., Baker, P. J. and Schneider, T. R. (2003) Optimization of selenium substructures as obtained from SHELXD. *Acta Cryst.* **D59,** 1987–1994.
58. Chen, C. C., Zhang, H., Kim, A. D., et al. (2002) Degradation pathway of the phosphonate ciliatine: crystal structure of 2-aminoethylphosphonate transaminase. *Biochemistry* **41,** 13,162–13,169.